


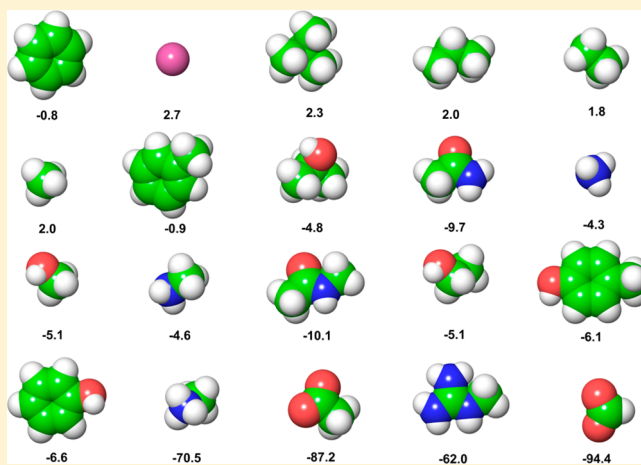
Estimating Translational and Orientational Entropies Using the *k*-Nearest Neighbors Algorithm

David J. Huggins*

Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, 19 J J Thomson Avenue, Cambridge CB3 0HE, United Kingdom

 Supporting Information

ABSTRACT: Inhomogeneous fluid solvation theory (IFST) and free energy perturbation (FEP) calculations were performed for a set of 20 solutes to compute the hydration free energies. We identify the weakness of histogram methods in computing the IFST hydration entropy by showing that previously employed histogram methods overestimate the translational and orientational entropies and thus underestimate their contribution to the free energy by a significant amount. Conversely, we demonstrate the accuracy of the *k*-nearest neighbors (KNN) algorithm in computing these translational and orientational entropies. Implementing the KNN algorithm within the IFST framework produces a powerful method that can be used to calculate free-energy changes for large perturbations. We introduce a new KNN approach to compute the total solute-water entropy with six degrees of freedom, as well as the translational and orientational contributions. However, results suggest that both the solute–water and water–water entropy terms are significant and must be included. When they are combined, the IFST and FEP hydration free energies are highly correlated, with an R^2 of 0.999 and a mean unsigned difference of 0.9 kcal/mol. IFST predictions are also highly correlated with experimental hydration free energies, with an R^2 of 0.997 and a mean unsigned error of 1.2 kcal/mol. In summary, the KNN algorithm is shown to yield accurate estimates of the combined translational-orientational entropy and the novel approach of combining distance metrics that is developed here could be extended to provide a powerful method for entropy estimation in numerous contexts.



INTRODUCTION

Inhomogeneous fluid solvation theory (IFST) is a statistical mechanical framework for calculating solvation free energies. The effect of a solute on the free energy of the surrounding solvent is quantified by calculating the enthalpy and entropy relative to bulk water.^{1–3} The solvation enthalpies are computed from interaction energies, and the solvation entropies are computed from intermolecular correlations. Only two-particle contributions are considered, and thus, the solvation entropy is calculated as the sum of a solute–water correlation term (S_{sw}) and a term from the change in water–water correlations (ΔS_{ww}). S_{sw} and ΔS_{ww} are expected to contribute unfavorably and favorably to the solvation free energy, respectively.⁴ IFST has been applied to proteins,^{5–7} peptides,⁸ host–guest complexes,⁹ small molecules,^{10,11} and bulk water.^{12–14} Previous work in this laboratory has demonstrated that hydration free energies for six small solutes calculated using IFST agree well with hydration free energies calculated using the more-established method of free-energy perturbation (FEP).¹⁰ In this case, the R^2 coefficient of determination was 0.99 and the mean unsigned difference

was 0.69 kcal/mol. However, this work also identified the problem with using a histogram method to calculate the IFST correlation functions. This problem is that the histogram bin sizes must be sufficiently small to capture the complexity of the probability density function but sufficiently large to avoid convergence issues. This is a well-known concern with histogram methods and has been addressed previously by a number of approaches.^{15,16} In the previous study, the amount of data from a 100 ns simulation was insufficient to yield converged entropy estimates for the required histogram bin sizes.¹⁰ The necessary use of inadequate histogram bin sizes led to underestimation of the entropy terms, as evidenced by the inability of the Cartesian coordinate system to recapitulate the radial distribution function of bulk water and the underestimate of the orientational entropy of bulk water. However, the S_{sw} and ΔS_{ww} terms are expected to be opposite in sign and the magnitudes of both are expected to be underestimated. Thus, a cancellation of errors is expected to yield a reasonable estimate

Received: May 13, 2014

of the solvation entropy, as observed. To avoid the inherent issue in histogram methods of selecting a bin size, other implementations of IFST have used the k -nearest neighbors (KNN) algorithm^{17–19} to calculate the orientational entropies.^{9,14} The KNN method yields an asymptotically unbiased entropy estimate, and we have recently demonstrated the accuracy of a KNN algorithm with an orientational distribution of known entropy.²⁰ In this work, we develop a KNN method to calculate the total solute–water solvation entropy S_{sw} . We then compute hydration free energies for a set of polar, nonpolar, and charged solutes using IFST and compare the results with hydration free energies calculated using FEP.²¹

METHODS

Hydration free energies were computed using FEP and IFST for the polar, nonpolar, and charged solutes listed in Table 1.

Table 1. Solute Studied In This Work^a

class	solute	initial RHDO size (Å)	water molecules	experimental ΔG hydration (kcal/mol)
nonpolar	benzene	25	376	−0.8 ⁴⁷
nonpolar	ethane	25	365	1.8 ⁴⁷
nonpolar	isobutane	25	377	2.3 ⁴⁷
nonpolar	methane	25	382	1.9 ⁴⁷
nonpolar	neon	25	358	2.7 ⁴⁷
nonpolar	propane	25	371	2.0 ⁴⁷
nonpolar	toluene	25	374	−0.9 ⁴⁷
polar	acetamide	25	384	−9.7 ⁴⁸
polar	ammonia	25	376	−4.3 ⁴⁷
polar	cresol	25	371	−6.1 ⁴⁹
polar	ethanol	25	364	−5.1 ⁴⁷
polar	methanol	25	374	−5.1 ⁴⁷
polar	methylamine	25	373	−4.6 ⁴⁷
polar	<i>N</i> -methyl acetamide	25	371	−10.1 ⁴⁸
polar	phenol	25	374	−6.6 ⁴⁹
polar	propan-2-ol	25	376	−4.8 ⁴⁹
charged	acetate	30	643	−78.7 ⁴⁴
charged	formate	30	631	−77.3 ⁴⁴
charged	methylammonium	30	651	−75.3 ⁴⁴
charged	methylguanidinium	30	631	−65.9 ⁴⁴

^aThe 20 solutes for which the hydration free energies were computed, along with the size of the RHDO periodic cell and the number of water molecules in the system for each case. The experimental hydration free energies and their sources are also noted.

The force field parameters for the molecules were taken from CHARMM36,²² and the force field parameters for neon were taken from CHARMM27.²³ The bond lengths, bond angles and dihedral angles were set to their force field equilibrium values for all molecules. Water molecules were modeled with the TIP4P-2005 water model.²⁴

Water Setup. A water shell of radius 50.0 Å was first generated around each solute with the SOLVATE program version 1.0 from the Max Planck Institute.²⁵ The resulting water globules were then cut to rhombic dodecahedral (RHDO) unit cells with side lengths of 25.0 Å for the uncharged solutes and 30.0 Å for the charged solutes. Larger systems were used for the charged solutes, as the perturbation to bulk water is expected to extend to the third solvation shell in this case.²⁶ To standardize the geometries of the water molecules, every hydrogen atom was deleted and all the necessary hydrogen atoms and lone pairs were built using the

appropriate geometry for TIP4P-2005 water. No additional ions were included in the systems.

IFST Protocol. Equilibration was performed for 1.0 ns in an NPT ensemble at 300 K and 1 atm using Langevin temperature control and Nosé–Hoover²⁷ Langevin piston pressure control.²⁸ All systems were brought to equilibrium before continuing, by verifying that the energy fluctuations were stable. MD simulations were performed using an MD time step of 2.0 fs. Electrostatic interactions were modeled with a uniform dielectric and a dielectric constant of 1.0 throughout the equilibration and production runs. van der Waals interactions were truncated at 11.0 Å with switching from 9.0 Å. Electrostatics were modeled using the particle mesh Ewald method,²⁹ and the systems were treated using rhombic dodecahedral periodic boundary conditions (PBC). All solute atoms were fixed for the entirety of the equilibration and production simulations. 80.0 ns of production simulation in an NPT ensemble were performed at 300 K and 1 atm for each system. System snapshots were saved every 400.0 fs, yielding 200 000 snapshots in total for each system. MD simulations were performed using NAMD³⁰ version 2.8 compiled for use with CUDA-accelerated GPUs.

IFST Calculations. IFST calculates the difference in free energy (ΔG_{IFST}) between a solution and the same number of solvent molecules (n) in the bulk, by combining the differences in interaction energy (ΔE_{IFST}) and entropy (ΔS_{IFST}).^{1,2} These are termed the local quantities and correspond to Ben-Naim's standard energy and entropy of solvation.³¹ If desired, ΔG_{IFST} can also be computed for small subvolumes, allowing the contribution of specific regions to be calculated and visualized.^{9–11} ΔE_{IFST} is calculated from the mean solute–water interaction energy (E_{sw}), the mean water–water interaction energy (E_{ww}), and the mean interaction energy of a bulk water molecule (E_{bulk}).

$$\begin{aligned}\Delta E_{\text{IFST}} &= E_{\text{sw}} + E_{\text{ww}} - nE_{\text{bulk}} \\ &= E_{\text{sw}} + \Delta E_{\text{ww}}\end{aligned}\quad (1)$$

E_{bulk} and E_{ww} are defined as half the interaction energy of a water molecule with all other water molecules in the system. We have implemented the minimum image convention to ensure consistent energy evaluations throughout the periodic cell.^{32–34} Previous work on quantitative application of IFST has shown that E_{bulk} must be calculated to high precision in order to yield accurate results.¹⁰ For the TIP4P-2005 water model and using the protocol described here, E_{bulk} is calculated from a 100 ns NPT simulation of 364 water molecules in a rhombic dodecahedral unit cell with side lengths of 25.0 Å at 300 K and 1 atm as −11.5702 kcal/mol. ΔS_{IFST} is calculated from the solute–water entropy (S_{sw}) and the difference in water–water entropy (ΔS_{ww}), with higher-order correlations not considered.

$$\begin{aligned}\Delta S_{\text{IFST}} &= S_{\text{sw}} + S_{\text{ww}} - nS_{\text{bulk}} \\ &= S_{\text{sw}} + \Delta S_{\text{ww}}\end{aligned}\quad (2)$$

The solute–water term (S_{sw}) is typically calculated as the sum of translational ($S_{\text{sw,trans}}$) and conditional orientational ($S_{\text{sw,orient}}$) contributions.

$$S_{\text{sw}} = S_{\text{sw,trans}} + S_{\text{sw,orient}}\quad (3)$$

It is important to note that IFST calculates entropies that are relative to a random distribution and are always negative (or zero for a random distribution). In this work, we develop novel

KNN approaches to calculate $S_{\text{sw,trans}}$ and S_{sw} . These can be combined to yield $S_{\text{sw,orient}}$ by substituting into eq 3. We calculate these quantities for the whole system rather than for a set of subvolumes. While it is valid to compute the contributions of each subvolume using this method, more data is required for proper convergence. We use a first nearest neighbor ($k = 1$) approach in all cases.¹⁷

$$S_{\text{sw,trans}} = nR \left\{ \frac{1}{nF} \sum_{i=1}^F \sum_{j=1}^n \ln \left[\frac{4\pi d_{\text{trans}}^3 nF}{3V_i} \right] + \gamma \right\} \quad (4)$$

R is the gas constant, F is the number of frames sampled, V_i is the volume of the system in frame i , and γ is Euler's constant, which corrects for the asymptotic bias. In this case the nearest neighbor distance (d_{trans}) is the Euclidean norm between the Cartesian coordinates of water molecule j in frame i , and its nearest neighbor water molecule k in frame l :

$$d_{\text{trans}} = \sqrt{(x_{ij} - x_{kl})^2 + (y_{ij} - y_{kl})^2 + (z_{ij} - z_{kl})^2} \quad (5)$$

For correct treatment of waters near the periodic boundary, the minimum image convention is used. The orientational distance (d_{orient}) between two water molecules is the distance between the rotations required to bring the two orientations to the same reference orientation. The correct distance metric for the rotation group is twice the geodesic distance on the unit sphere.²⁰

$$d_{\text{orient}} = \|\log(\mathbf{R}_{ij}\mathbf{R}_{kl}^T)\| \\ = 2 \times \text{acos}(|\mathbf{q}_{ij} \cdot \mathbf{q}_{kl}|) \quad (6)$$

This matrix representation of the rotations for water molecule j in frame i and its nearest neighbor water molecule k in frame l are denoted by \mathbf{R}_{ij} and \mathbf{R}_{kl} and the quaternion representations are denoted by \mathbf{q}_{ij} and \mathbf{q}_{kl} . $\|\mathbf{M}\|$ and \mathbf{M}^T represent the Euclidean (Frobenius) norm and the transpose of the matrix \mathbf{M} respectively. The total solute–water entropy is estimated by combining the translational and orientational distance metrics.

$$S_{\text{sw}} = nR \left\{ \frac{1}{nF} \sum_{i=1}^F \sum_{j=1}^n \ln \left[\frac{\pi d_{\text{total}}^6 nF}{48V_i} \right] + \gamma \right\} \quad (7)$$

$$d_{\text{total}} = \sqrt{d_{\text{trans}}^2 + d_{\text{orient}}^2} \quad (8)$$

We noted from previous work that the ratio of E_{sw} and ΔE_{ww} was approximately minus a half, as shown in Table 2.¹⁰ Thus, solute–water interactions are offset by reduced water–water interactions. Prompted by this observation, ΔS_{ww} is calculated in a very simple manner in this work.

Table 2. Observed Relationship Between E_{sw} and ΔE_{ww} ^a

solute	E_{sw} (kcal/mol)	ΔE_{ww} (kcal/mol)	$\Delta E_{\text{ww}}/E_{\text{sw}}$
acetamide	−29.44	14.78	−0.50
benzene	−15.78	8.33	−0.53
isobutane	−9.83	4.70	−0.48
methane	−3.64	1.17	−0.32
methanol	−19.68	9.73	−0.49
<i>N</i> -methyl acetamide	−29.10	14.22	−0.49

^aThe solute–water interaction energy (E_{sw}) and the change in water–water interaction energy (ΔE_{ww}) for the six solutes studied previously.

$$\Delta S_{\text{ww}} = -1/2S_{\text{sw}} \quad (9)$$

Thus, solute–water correlations are offset by reduced water–water correlations. Once ΔE_{IFST} and ΔS_{IFST} have been calculated, they can then be combined to yield ΔG_{IFST} .

$$\Delta G_{\text{IFST}} = \Delta E_{\text{IFST}} - T\Delta S_{\text{IFST}} \quad (10)$$

FEP Protocol. The equilibrated systems for the IFST simulations were used as the startpoints for the FEP systems. These systems consist of the solute in water and correspond to the $\lambda=0.0$ states. FEP calculations were performed in both forward and backward directions to yield corresponding predictions for annihilation ($\lambda=0.0$ to $\lambda=1.0$) and creation ($\lambda=1.0$ to $\lambda=0.0$) of the solutes. Each annihilation and creation was split into 24 steps to yield 48 λ windows per system. The lambda schedules for the uncharged and charged solutes are reported in Tables S1 and S2 respectively. Two measures were adopted to avoid the numerical instabilities that occur when λ approaches 0.0 or 1.0. First a soft-core potential was employed with a van der Waals radius-shifting coefficient of 5.0 for the uncharged solutes and 4.0 for the charged solutes.^{35,36} Second, electrostatic interactions were scaled down to zero between $\lambda=0.0$ and $\lambda=0.4$ for uncharged solutes and between $\lambda=0.0$ and $\lambda=0.575$ for charged solutes.³⁷ Starting with the equilibrated systems generated for the IFST simulations, further equilibration was performed at 300 K and 1 atm for 250 ps in an NPT ensemble for each lambda window. This was followed by 1.0 ns of production simulation in an NPT ensemble. MD simulations were performed using NAMD³⁰ version 2.8.

FEP Calculations. The change in free energy (ΔG_{FEP}) was calculated as the sum of free energy changes for a series of N small steps between intermediate states a and b .²¹

$$\Delta G_{\text{FEP}} = \sum_{a=1, b=a+1}^N \Delta G_{a \rightarrow b} \quad (11)$$

The change in free energy was calculated for each small step ($\Delta G_{a \rightarrow b}$) using the partition functions (Q) for the two states, which are calculated from the Hamiltonians (H).

$$\Delta G_{a \rightarrow b} = G_b - G_a = -kT \ln \left(\frac{Q_b}{Q_a} \right) \\ = -kT \ln \langle \exp(-(H_b - H_a)/kT) \rangle_a \quad (12)$$

The results for the forward and backward FEP simulations were combined using the Bennett Acceptance Ratio (BAR) method.^{37,38} BAR was implemented using the ParseFEP Plugin from VMD.³⁹ In the context of FEP, the solvation energy (ΔE_{FEP}) is simply the difference in total interaction energy between the solution and the bulk solvent. It is thus equal to the total solvation energy from IFST.

$$\Delta E_{\text{FEP}} = \Delta E_{\text{IFST}} \quad (13)$$

The estimated statistical error in the FEP free energy predictions using BAR was less than 0.5 kcal/mol in all cases.

Additional Considerations. As discussed previously, IFST ignores the small nonlocal contributions to the solvation free energy¹ and the small contribution of the volume change to the solvation enthalpy ($P\Delta V$).¹⁰ In addition to this, using a nonpolarizable force field means that the free energy changes associated with polarization of the solutes and the water molecules are ignored for both FEP and IFST.⁴⁰

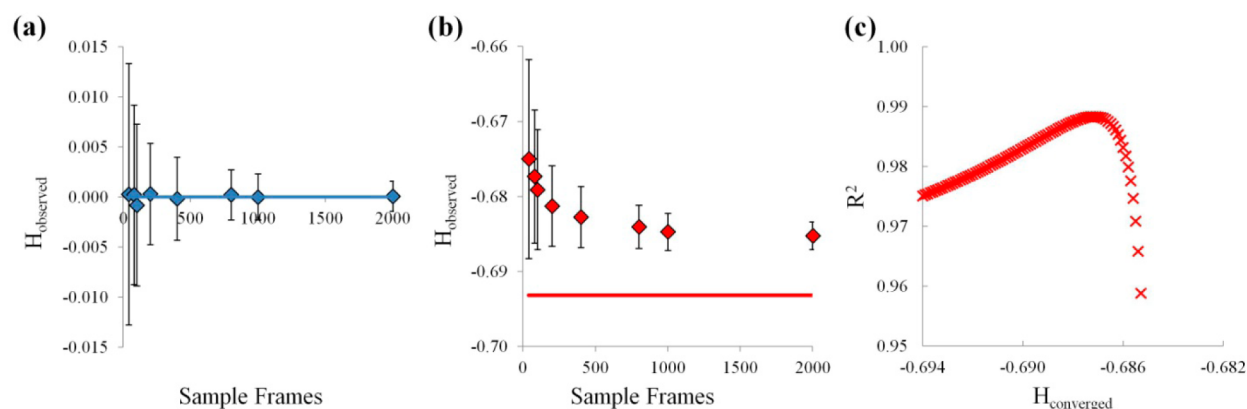


Figure 1. Convergence of the calculated translational entropy. Convergence of the calculated relative translational entropy with increased sampling for the randomly generated uniform data (a) and biased data (b). The expected relative translational entropies are marked as solid lines. The plotted points represent 40, 80, 100, 200, 400, 800, 1000, and 2000 sample frames in each case and the error bars represent the standard deviations for each level of sampling. (c) The coefficient of determination between $\ln[H_{\text{observed}} - H_{\text{converged}}]$ and $\ln(N)$ for the biased data at 100 equally spaced values of $H_{\text{converged}}$ between -0.6583 and -0.6952 .

One very important point must also be considered for the four charged solutes. For systems with non-zero charge, the use of PBC requires correction terms to calculate thermodynamic properties, such as the solvation free energy, correctly.^{41–43} In this work, we use correction terms used previously to study charged polyatomic solutes.⁴⁴ We computed the ΔG_{A+B+D} and ΔG_{Cl} correction terms using equations 14 and 15, respectively.

$$\Delta G_{A+B+D} = \Delta G_{\text{Non-PBC}}^{\text{PB}} - \Delta G_{\text{PBC}}^{\text{PB}} \quad (14)$$

$$\Delta G_{\text{Cl}} = - \frac{N_A \rho_w \gamma_w q}{6 \epsilon_0} \quad (15)$$

The terms on the right-hand side of equation 14 are charging energies calculated using the Poisson–Boltzmann (PB) solver in CHARMM⁴⁵ using non-PBC and PBC with dielectric constants of 78.4 (experimental) and 60 (TIP4P-2005) respectively. In equation 15, N_A is Avogadro's constant, ρ_w is the average number density of water molecules in the system, γ_w is the quadrupole moment trace of the solvent model (0.0099 e-nm² for TIP4P-2005), q is the formal system charge, and ϵ_0 is the vacuum permittivity. The hydration free energies for the charged solutes calculated using FEP and IFST were then corrected by these two terms. We did not calculate the separate correction terms for the enthalpy and entropy.

Analysis of Results. The results from FEP and IFST are compared by two methods. The correlation between the predictions is assessed by computing the coefficient of determination (R^2), and the difference in prediction is assessed by computing the mean unsigned difference (MUD).

Randomly Generated Data. The KNN approach to calculating $S_{\text{sw,trans}}$ and S_{sw} was first assessed using randomly generated data. Relevant data were produced by generating Cartesian coordinates and orientations for 364 molecules in a cubic box. The edge length of the box was randomly varied between 99.0% and 101.0% of 25.0 Å to mimic the NPT ensemble that we are interested in. This process was repeated to generate separate frames and $S_{\text{sw,trans}}$ and S_{sw} were calculated using eqs 4 and 7, respectively. This process was repeated 96 times to calculate a mean and standard deviation. To generate biased data with known entropy, the Cartesian coordinates x , y , and z can be divided by the divisors A , B , and C , and the orientations can be restricted to within a certain distance of the

reference orientation by the divisor D .^{20,46} The expected relative entropy (H_{biased}) can then be calculated using eq 16.

$$H_{\text{biased}} = \ln[\pi/D - \sin \pi/D] - \ln(\pi) - \ln(ABC) \quad (16)$$

RESULTS AND DISCUSSION

The first stage of analysis was to validate the new approach to calculating the relative translational entropy using eq 4. We used different numbers of randomly generated samples to explore the convergence properties of the calculation and the results are presented in Figure 1.

As expected, increased sampling leads to more accurate entropy predictions and reduced standard deviations. For the biased data, 200 sample frames yield a mean predicted entropy that is within 2.0% of the expected entropy. It is interesting to note that the observed entropies (H_{observed}) in Figure 1b are well approximated by a power law that approaches a converged entropy ($H_{\text{converged}}$). For a given number of samples (N) and constants k and p , this can be expressed as follows:

$$H_{\text{observed}} = H_{\text{converged}} + kN^{-p} \quad (17)$$

$$\ln[H_{\text{observed}} - H_{\text{converged}}] = \ln(k) - p \ln(N) \quad (18)$$

This allows $H_{\text{converged}}$ to be estimated using a series of H_{observed} from different number of samples, by finding the value of $H_{\text{converged}}$ that maximizes the correlation between $\ln[H_{\text{observed}} - H_{\text{converged}}]$ and $\ln(N)$. Figure 1c shows a plot of this correlation for 100 estimates of $H_{\text{converged}}$. The maximum value of R^2 corresponds to the value of $H_{\text{converged}}$ that best fits the data. This observation has been noted in previous work using the KNN approach and can be a useful technique in estimating entropies using limited data.¹⁸ However, we do not use this approach to compute the thermodynamic entropies in this work. For the randomly generated biased data in this case, the optimal value of $H_{\text{converged}}$ is -0.6872 and the expected relative entropy is -0.6931 . We further investigated the accuracy of the method by calculating the relative translational entropy for increasingly biased data. Figure 2 illustrates a plot of the expected and predicted entropies for four cases.

The standard deviations are less than 0.003 entropy units in each case. This corresponds to a thermodynamic entropy of approximately 0.005 cal/K-mol for each particle. After assessing

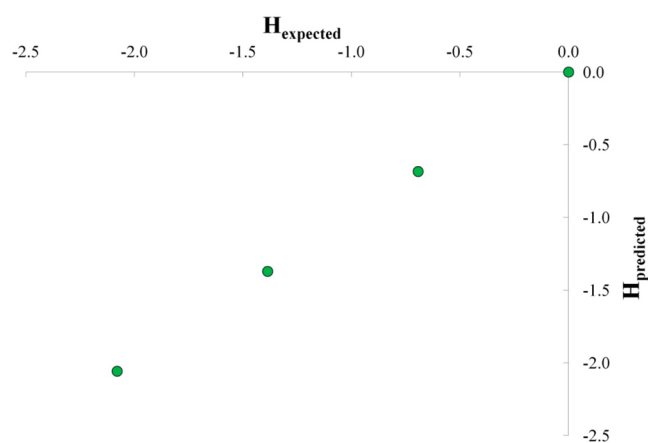


Figure 2. Predicted and expected entropies for increasingly biased data. Observed and expected translational entropies for the cases $A = 1/B = 1/C = 1$, $A = 1/B = 1/C = 2$, $A = 1/B = 2/C = 2$, and $A = 2/B = 2/C = 2$. The expected entropies are calculated using eq 16 with $D = 1$.

the precision and accuracy of the KNN method, we moved on to study data from MD simulations. As shown previously, KNN only yields accurate estimates of the entropy if the samples are independent.²⁰ For the conditional relative orientational entropy in bulk water it was demonstrated that a sampling interval of greater than 400 fs was necessary. We explored the sampling interval in the context of relative translational entropy for simulations of bulk water and a solution of benzene. Figure 3a shows the convergence of $-T\Delta S_{sw,trans}$ for increasing sampling intervals with a constant number of frames (2000) and Figure 3b shows the convergence of $-T\Delta S_{sw,trans}$ for increasing simulation time with a constant time step (20000 fs).

The results demonstrate that a sampling interval of 400 fs does not yield independent samples for the relative translational entropy. In this case, a sampling interval of greater than 1 ps is necessary. However, they also suggest that an acceptable degree of convergence is provided by relatively few frames. After

reviewing this data, we employed a sampling interval of 40.0 ps and a simulation time of 80.0 ns to calculate the relative translational entropy of the 20 solutes, yielding 2000 frames.

After considering the relative translational entropy, we repeated the analysis to validate the new approach to calculating the total solute-water entropy using eq 7. The results from analyzing the convergence using random data are presented in Figure 4.

As was the case for the relative translational entropy, the convergence is well approximated by a power law, with an optimal value of $H_{converged}$ at -2.364 while the expected relative entropy is -2.399 . Again, increased sampling lead to more accurate entropy predictions and 100 sample frames yield a mean predicted entropy that is within 10.0% of the expected entropy for the biased data. The accuracy of the combined translational-orientational entropy estimates are illustrated in Figure 5, which is a plot of the expected and predicted entropies for five cases.

We explored the sampling interval in the context of combined translational-orientational entropy for simulations of bulk water and a solution of benzene. Figure 6a shows the convergence of $-T\Delta S_{sw}$ for increasing sampling intervals with a constant number of frames (20 000) and Figure 3b shows the convergence of $-T\Delta S_{sw}$ for increasing simulation time with a constant time step (4000 fs).

The results demonstrate that a sampling interval of greater than 2.0 ps is necessary for the relative combined translational-orientational entropy but that acceptable convergence is again reached with relatively few frames. We employed a sampling interval of 4.0 ps and a simulation time of 80.0 ns to calculate the relative combined translational-orientational entropy of the 20 solutes, yielding 20000 frames. Having determined the sampling interval and simulation time necessary for sufficient convergence, we moved on to consider the 20 solutes. Table 3 reports the thermodynamic predictions for the 20 solutes calculated using FEP and IFST. If one computes the hydration free energy as the sum of ΔE_{IFST} and $-T\Delta S_{sw}$ (ignoring the $-T\Delta S_{ww}$ term) the correlation with ΔG_{FEP} is very good ($R^2 =$

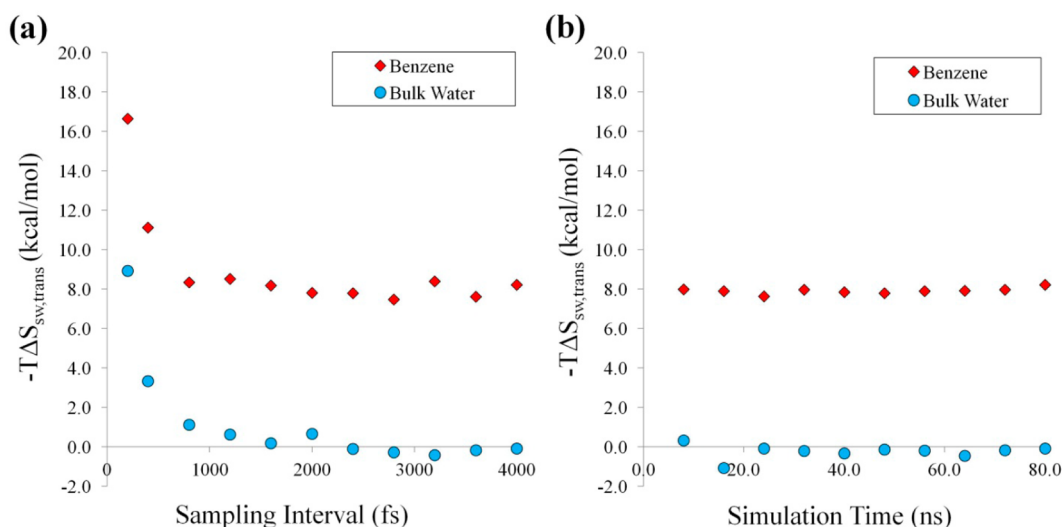


Figure 3. Effect of sampling interval and sampling time on $-T\Delta S_{sw,trans}$. (a) Effect of increasing the sampling interval for a fixed number of 2000 frames on the relative translational entropy for bulk water (blue circles) and benzene (red diamonds). The plotted points represent 200, 400, 800, 1200, 1600, 2000, 2400, 2800, 3200, 3600, and 4000 fs. (b) Effect of increasing the simulation time for a fixed sampling interval of 20 000 fs on the relative translational entropy for bulk water (blue circles) and benzene (red diamonds). The plotted points represent 8.0, 16.0, 24.0, 32.0, 40.0, 48.0, 56.0, 64.0, 72.0, and 80.0 ns.

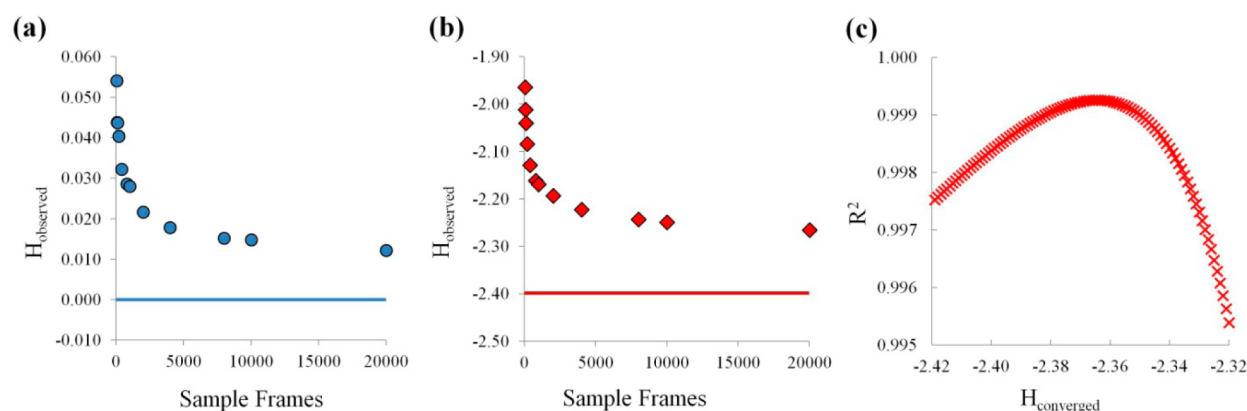


Figure 4. Convergence of the combined translational-orientational entropies. Convergence of the calculated relative entropy with increased sampling for the randomly generated uniform data (a) and biased data (b). The plotted points represent 40, 80, 100, 200, 400, 800, 1000, 2000, 4000, 8000, 10000, and 20000 frames in each case. (c) The coefficient of determination between $\ln[H_{\text{observed}} - H_{\text{converged}}]$ and $\ln(N)$ for 100 equally spaced values of $H_{\text{converged}}$ between -2.32 and -2.42 .

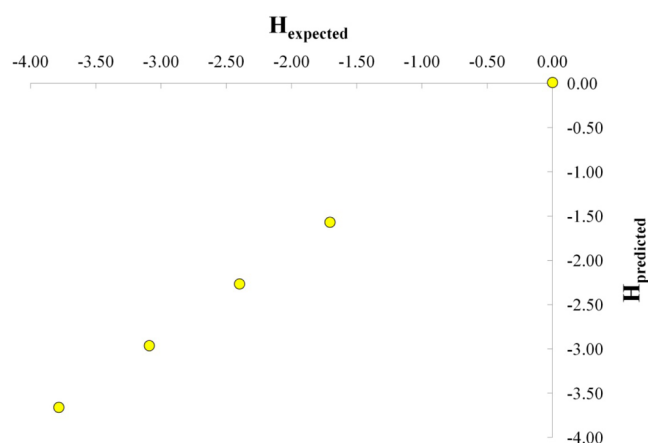


Figure 5. Observed and expected entropies for increasingly biased data. Observed and expected entropies for the cases $A = 1/B = 1/C = 1/D = 1$, $A = 1/B = 1/C = 2/D = 1$, $A = 1/B = 2/C = 2/D = 1$, $A = 2/B = 2/C = 2/D = 1$, and $A = 2/B = 2/C = 2/D = 2$. The expected entropies are calculated using eq 16.

0.992), but the mean unsigned difference (MUD) is 5.6 kcal/mol. If one includes the $-T\Delta S_{\text{sw}}$ term, calculated using eq 9, the correlation between ΔG_{IFST} and ΔG_{FEP} is still very good ($R^2 = 0.999$) and the MUD is 0.9 kcal/mol. This result illustrates the importance of including the $-T\Delta S_{\text{sw}}$ term. The excellent correlation can be seen in Figure 7, which shows the FEP and IFST predictions of hydration free energy for the 20 solutes.

We also assessed the ability of IFST to accurately predict the experimental hydration properties. Figure 8 shows the correlation between the IFST predictions and the experimental hydration free energies.

The correlation between ΔG_{IFST} and $\Delta G_{\text{Experimental}}$ is excellent ($R^2 = 0.997$), with a mean unsigned error (MUE) of 1.2 kcal/mol. These results suggest that the TIP4P-2005 water model in combination with the CHARMM forcefield is suitable for quantitative application using IFST. However, future applications of IFST should consider using KNN entropy estimates in place of histogram entropy estimates. Table 4 reports the translational, orientational and total entropies for six solutes calculated IFST with a histogram method and a KNN method. The data shows that the histogram method underestimates

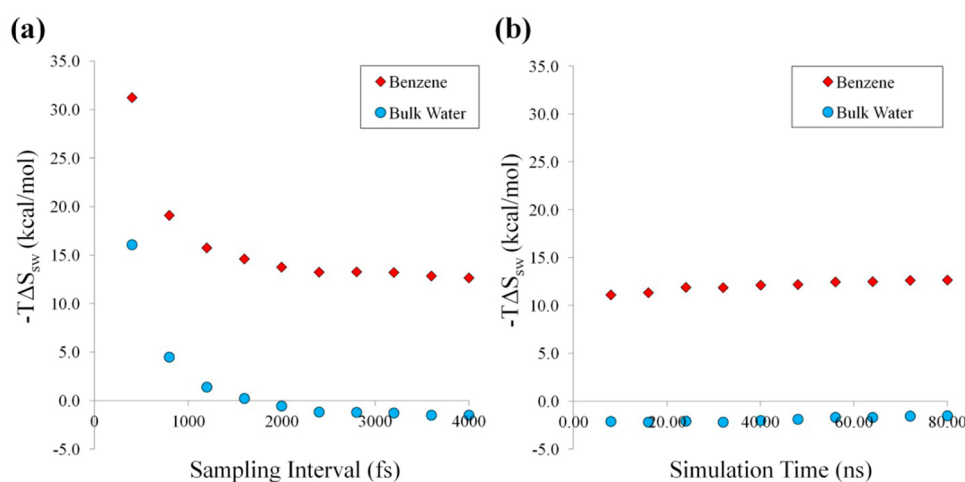


Figure 6. Effect of sampling interval and sampling time on $-T\Delta S_{\text{sw}}$. (a) Effect of increasing the sampling interval for a fixed number of 20 000 frames on the relative entropy for bulk water (blue circles) and benzene (red diamonds). The plotted points represent 800, 1200, 1600, 2000, 2400, 2800, 3200, 3600, and 4000 fs. (b) Effect of increasing the simulation time for a fixed sampling interval of 4000 fs on the relative translational entropy for bulk water (blue circles) and benzene (red diamonds). The plotted points represent 8.0, 16.0, 24.0, 32.0, 40.0, 48.0, 56.0, 64.0, 72.0, and 80.0 ns.

Table 3. Thermodynamic Predictions from IFST and FEP^a

class	solute	ΔE (kcal/mol)	$-T\Delta S_{sw}$ (kcal/mol)	$-T\Delta S_{ww}$ (kcal/mol)	ΔG_{IFST} (kcal/mol)	ΔG_{FEP} (kcal/mol)
nonpolar	benzene	-6.7	12.6	-6.3	-0.4	0.4
nonpolar	ethane	-2.9	7.7	-3.9	1.0	2.5
nonpolar	isobutane	-4.0	11.8	-5.9	1.9	3.1
nonpolar	methane	-2.6	5.3	-2.6	0.0	2.3
nonpolar	neon	0.7	2.2	-1.1	1.9	2.1
nonpolar	propane	-3.4	9.9	-5.0	1.6	2.8
nonpolar	toluene	-7.7	14.8	-7.4	-0.3	0.8
polar	acetamide	-13.5	11.2	-5.6	-7.9	-8.3
polar	ammonia	-6.6	5.9	-3.0	-3.7	-3.7
polar	cresol	-12.7	17.8	-8.9	-3.8	-4.4
polar	ethanol	-11.7	12.4	-6.2	-5.5	-4.9
polar	methanol	-9.8	8.9	-4.4	-5.4	-4.6
polar	methylamine	-9.0	9.3	-4.6	-4.3	-4.2
polar	N-methyl acetamide	-13.9	13.8	-6.9	-7.0	-6.8
polar	phenol	-11.3	14.9	-7.4	-3.8	-3.7
polar	propan-2-ol	-12.6	15.5	-7.7	-4.8	-4.2
charged	acetate	-95.8	24.6	-12.3	-77.6	-74.5
charged	formate	-95.5	20.5	-10.3	-78.2	-76.4
charged	methylammonium	-45.5	11.9	-5.9	-75.3	-76.3
charged	methylguanidinium	-40.7	18.0	-9.0	-66.9	-68.4

^aThe IFST predictions of ΔE , $-T\Delta S_{sw}$, and $-T\Delta S_{ww}$ and the comparison of ΔG_{IFST} and ΔG_{FEP} for the 20 solutes studied. The corrections for using PBC have been applied to ΔG_{IFST} and ΔG_{FEP} for the charged systems.

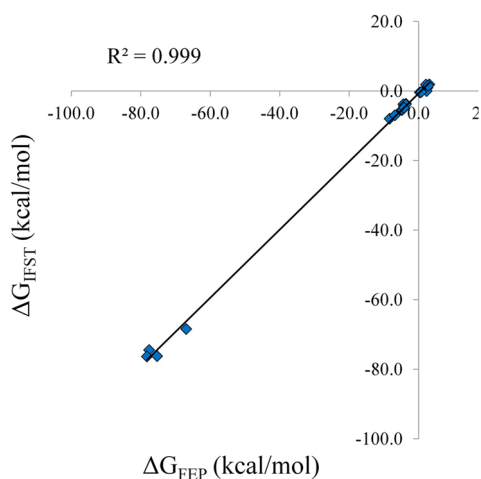


Figure 7. The predictions of ΔG from FEP and IFST. The slope of the trendline is 0.98.

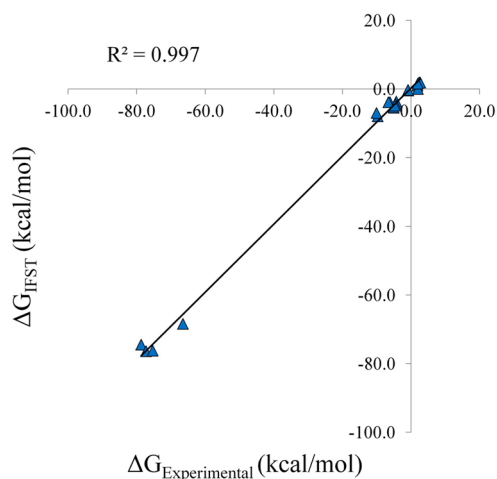


Figure 8. The predictions of ΔG from IFST plotted against the experimental quantities. The slope of the trendline is 0.99.

$-T\Delta S_{sw,trans}$ by approximately 2-fold and $-T\Delta S_{sw,orient}$ by approximately 3-fold for the bin sizes used previously.¹⁰ Thus, ignoring or underestimating $-T\Delta S_{ww}$ can lead to approximate cancellation of errors for certain cases. However, if the $-T\Delta S_{sw}$ term is calculated accurately, then the $-T\Delta S_{ww}$ term becomes important, as shown by the MUD of 5.6 kcal/mol when it is excluded and the MUD of 0.9 kcal/mol when it is included.

CONCLUSIONS

This study addresses the quantitative accuracy of IFST by comparing hydration free energies from IFST with hydration free energies from FEP. In particular, we compared the estimates of solute-water entropy using the KNN algorithm with estimates from a histogram method. Using randomly generated data, we demonstrated that the KNN algorithm yields accurate predictions of translational entropy and

combined translational-orientational entropy. We note that the KNN entropy estimates approach a converged entropy as more samples are used and that this convergence is well approximated by a power law. This has been noted previously and can be used to improve entropy estimates.¹⁸ Comparison of KNN and histogram estimates for data from MD suggests that the histogram method overestimates the entropy and thus underestimates $-T\Delta S$ to a significant degree. For a grid resolution of 0.5 Å and an angular bin size was 45°, the histogram method underestimates $-T\Delta S_{sw,trans}$ by approximately 2-fold and $-T\Delta S_{sw,orient}$ by approximately 3-fold. This problem is masked to some degree if a histogram method is also employed to calculate $-T\Delta S_{ww}$ due to cancellation of errors. However, if a KNN method is used to calculate $-T\Delta S_{sw}$ then $-T\Delta S_{ww}$ is predicted to be significant in magnitude and must be included. In short, it is clear that histogram methods are not suitable for quantitative applications of IFST with the

Table 4. Comparisons of Entropy Estimates from KNN and Histogram Methods^a

solute	histogram $-T\Delta S_{sw}$ (kcal/mol)	KNN $-T\Delta S_{sw}$ (kcal/mol)	histogram $-T\Delta S_{trans}$ (kcal/mol)	KNN $-T\Delta S_{trans}$ (kcal/mol)	histogram $-T\Delta S_{orient}$ (kcal/mol)	KNN $-T\Delta S_{orient}$ (kcal/mol)
acetamide	6.6	14.8	3.7	6.7	2.9	8.1
benzene	7.4	16.6	4.7	8.4	2.7	8.1
isobutane	7.0	15.0	4.6	7.6	2.4	7.4
methane	3.9	9.0	2.3	4.5	1.6	4.4
methanol	5.6	12.4	3.1	5.6	2.5	6.8
N-methyl acetamide	7.7	17.2	4.5	7.8	3.2	9.3

^aThe total solute–water entropies, translational solute–water entropies, and orientational solute–water entropies for six solutes calculated using the KNN algorithm and using a grid-based histogram method. For the histogram method, the grid resolution was 0.5 Å and the angular bin size was 45°.

bin sizes used previously. While histogram methods could be used to calculate accurate entropies by using smaller bin sizes, this would introduce impractical sampling requirements.

The results of this study also reinforce the finding that short MD sampling intervals yield correlated samples, which leads to skewed KNN entropy estimates.²⁰ For accurate estimation of the translational-orientational entropy, a sampling interval of greater than 2.0 ps is necessary to yield sufficiently uncorrelated samples. It is probable that this problem could be fixed by significantly increased sampling, but this is neither feasible nor desirable. The extremely good agreement with hydration free energies from FEP ($R^2 = 0.999$ and MUD = 0.9 kcal/mol) suggest that IFST is a useful free-energy method with the major advantage that it deals equally well with large and small perturbations. Comparison with experimental hydration free energies also demonstrates that the TIP4P-2005 water model is suitable for quantitative applications of IFST ($R^2 = 0.997$ and MUE = 1.2 kcal/mol). In addition, the success of the KNN algorithm in estimating the combined translational-orientational entropy suggests that a similar method could be used to estimate entropies for other correlated degrees of freedom, provided that the correct distance metrics can be identified. This could provide a powerful and general method for entropy estimation in numerous contexts.

■ ASSOCIATED CONTENT

● Supporting Information

The FEP lambda schedules for the uncharged and charged solutes are reported in Tables S1 and S2, respectively. This information is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: djh210@cam.ac.uk. Telephone: +44(0)1223763367.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by the MRC under grant ML/L007266/1. All calculations were performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (<http://www.hpc.cam.ac.uk/>) provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and were funded by the EPSRC under grants EP/F032773/1 and EP/J017639/1.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Acknowledgements go to Stuart Rankin for technical help and Mike Gilson, David Mobley, Philippe Hünenberger, and Maria Reif for helpful discussions.

■ REFERENCES

- Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J. Phys. Chem. B* **1998**, *102* (18), 3531–3541.
- Lazaridis, T. Inhomogeneous fluid approach to solvation thermodynamics. 2. Applications to simple fluids. *J. Phys. Chem. B* **1998**, *102* (18), 3542–3550.
- Li, Z.; Lazaridis, T. Computing the thermodynamic contributions of interfacial water. *Methods Mol. Biol.* **2012**, *819*, 393–404.
- Huggins, D. J. Benchmarking the thermodynamic analysis of water molecules around a model beta sheet. *J. Comput. Chem.* **2012**, *33* (15), 1383–1392.
- Li, Z.; Lazaridis, T. Thermodynamics of buried water clusters at a protein–ligand binding interface. *J. Phys. Chem. B* **2006**, *110* (3), 1464–1475.
- Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein–ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (3), 808–813.
- Huggins, D. J.; Marsh, M.; Payne, M. C. Thermodynamic Properties of Water Molecules at a Protein Protein Interaction Surface. *J. Chem. Theory Comput.* **2011**, *7* (11), 3514–3522.
- Czapiewski, D.; Zielkiewicz, J. Structural Properties of Hydration Shell Around Various Conformations of Simple Polypeptides. *J. Phys. Chem. B* **2010**, *114* (13), 4536–4550.
- Nguyen, C. N.; Young, T. K.; Gilson, M. K. Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137* (4), 044101.
- Huggins, D. J.; Payne, M. C. Assessing the Accuracy of Inhomogeneous Fluid Solvation Theory in Predicting Hydration Free Energies of Simple Solutes. *J. Phys. Chem. B* **2013**, *117* (27), 8232–8244.
- Lazaridis, T. Solvent reorganization energy and entropy in hydrophobic hydration. *J. Phys. Chem. B* **2000**, *104* (20), 4964–4979.
- Lazaridis, T.; Karplus, M. Orientational correlations and entropy in liquid water. *J. Chem. Phys.* **1996**, *105* (10), 4294–4316.
- Esposito, R.; Saija, F.; Saitta, A. M.; Giaquinta, P. V. Entropy-based measure of structural order in water. *Phys. Rev. E* **2006**, *73* (4), 040502.
- Wang, L.; Abel, R.; Friesner, R. A.; Berne, B. J. Thermodynamic properties of liquid water: an application of a nonparametric approach to computing the entropy of a neat fluid. *J. Chem. Theory Comput.* **2009**, *5* (6), 1462–1473.
- Herzel, H.; Schmitt, A.; Ebeling, W. Finite sample effects in sequence analysis. *Chaos, Solitons Fractals* **1994**, *4* (1), 97–113.

- (16) Numata, J.; Knapp, E.-W. Balanced and bias-corrected computation of conformational entropy differences for molecular trajectories. *J. Chem. Theory Comput.* **2012**, *8* (4), 1235–1245.
- (17) Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.* **2003**, *23* (3–4), 301–321.
- (18) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.* **2007**, *28* (3), 655–668.
- (19) Hnizdo, V.; Tan, J.; Killian, B. J.; Gilson, M. K. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comput. Chem.* **2008**, *29* (10), 1605–1614.
- (20) Huggins, D. J. Comparing distance metrics for rotation using the k-nearest neighbors algorithm for entropy estimation. *J. Comput. Chem.* **2014**, *35* (5), 377–385.
- (21) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420.
- (22) Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; MacKerell, A. D., Jr.; Pastor, R. W. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **2010**, *114* (23), 7830–7843.
- (23) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.
- (24) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **2005**, *123* (23), 234505.
- (25) Grubmüller, H. Groll, V. *Solvate: A Program to Create Atomic Solvent Models*, Version 1.0.1; University of Munich, 1996.
- (26) Agarwal, M.; Kushwaha, H. R.; Chakravarty, C. Local Order, Energy, and Mobility of Water Molecules in the Hydration Shell of Small Peptides. *J. Phys. Chem. B* **2010**, *114* (1), 651–659.
- (27) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant-Pressure Molecular-Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101* (5), 4177–4189.
- (28) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method. *J. Chem. Phys.* **1995**, *103* (11), 4613–4621.
- (29) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103* (19), 8577–8593.
- (30) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.
- (31) Ben-Naim, A. Standard thermodynamics of transfer. Uses and misuses. *J. Chem. Phys.* **1978**, *82* (7), 792–803.
- (32) Smith, W. The periodic boundary condition in non-cubic MD cells: Wigner-Seitz cells with reflection symmetry. *CCPS Inf. Q. Comput. Simul. Condens. Phases, Informal Newslett.* **1983**, *10*, 37.
- (33) Adams, D. Alternatives to the periodic cube in computer simulation. *CCPS Inf. Q. Comput. Simul. Condens. Phases, Informal Newslett.* **1983**, *10*, 30–36.
- (34) Smith, W. The minimum image convention in non-cubic MD cells. *CCPS Inf. Q. Comput. Simul. Condens. Phases, Informal Newslett.* **1989**, *30*, 35.
- (35) Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding Singularities and Numerical Instabilities in Free-Energy Calculations Based on Molecular Simulations. *Chem. Phys. Lett.* **1994**, *222* (6), 529–539.
- (36) Zacharias, M.; Straatsma, T. P.; Mccammon, J. A. Separation-Shifted Scaling, a New Scaling Method for Lennard-Jones Interactions in Thermodynamic Integration. *J. Chem. Phys.* **1994**, *100* (12), 9025–9031.
- (37) Pohorille, A.; Jarzynski, C.; Chipot, C. Good practices in free-energy calculations. *J. Phys. Chem. B* **2010**, *114* (32), 10235–10253.
- (38) Bennett, C. H. Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J. Comput. Phys.* **1976**, *22* (2), 245–268.
- (39) Liu, P.; Dehez, F.; Cai, W. S.; Chipot, C. A Toolkit for the Analysis of Free-Energy Perturbation Calculations. *J. Chem. Theory Comput.* **2012**, *8* (8), 2606–2616.
- (40) Hess, B.; van der Vegt, N. F. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J. Phys. Chem. B* **2006**, *110* (35), 17616–17626.
- (41) Ashbaugh, H. S.; Wood, R. H. Effects of long-range electrostatic potential truncation on the free energy of ionic hydration. *J. Chem. Phys.* **1997**, *106* (19), 8135–8139.
- (42) Reif, M. M.; Hünenberger, P. H. Computation of methodology-independent single-ion solvation properties from molecular simulations. III. Correction terms for the solvation free energies, enthalpies, entropies, heat capacities, volumes, compressibilities, and expansivities of solvated ions. *J. Chem. Phys.* **2011**, *134* (14), 144103.
- (43) Rocklin, G. J.; Mobley, D. L.; Dill, K. A.; Hünenberger, P. H. Calculating the binding free energies of charged species based on explicit-solvent simulations employing lattice-sum methods: An accurate correction scheme for electrostatic finite-size effects. *J. Chem. Phys.* **2013**, *139* (18), 184103.
- (44) Reif, M. M.; Hünenberger, P. H.; Oostenbrink, C. New interaction parameters for charged amino acid side chains in the GROMOS force field. *J. Chem. Theory Comput.* **2012**, *8* (10), 3705–3723.
- (45) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ochinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.
- (46) Li, S. Concise formulas for the area and volume of a hyperspherical cap. *Asian J. Math. Stat.* **2011**, *4* (1), 66–70.
- (47) Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Chem. Phys.* **1984**, *81* (4), 2016–2027.
- (48) Wolfenden, R. Interaction of the peptide bond with solvent water: a vapor phase analysis. *Biochemistry* **1978**, *17* (1), 201–204.
- (49) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group contributions to the thermodynamic properties of non-ionic organic solutes in dilute aqueous solution. *J. Solution Chem.* **1981**, *10* (8), 563–595.